University of Ljubljana

Doctoral Programme in Statistics

Methodology of Statistical Research

Written examination

February 12th, 2021

Name and surname: ________________ ID number: | | | | | | | | |

Instructions

Read carefully the wording of the problem before you start. There are four problems altogether. You may use a A4 sheet of paper and a mathematical handbook. Please write all the answers on the sheets provided. You have two hours.

Problem	a.	b.	c.	d.	
1.				•	
2.					
3.			•	•	
4.				•	
Total					

1. (25) Often it is difficult to obtain honest answers from sample subjects to questions like "Have you ever used heroin" or "Have you ever cheated on an exam". To reduce bias the method or *randomized response* is used. The sample subject is given one of the two statements below at random:

(1) "I have property A.'

(2) "I do not have property A.'

The subject responds YES or NO to the given question. The pollster does not know to which of the two statements the subject is responding. We assume:

- The subjects are a simple random sample of size n from a larger population of size N.
- The statements are assigned to the chosen subjects independently.
- The assignment of statements is independent of the sampling procedure.
- The subjects respond honestly to the statements they are given.

Let

- p be the probability the a subject will be assigned the statement (1). This probability is known and is part of the design.
- q be the proportion of subjects in the population with property A.
- r be the probability that a randomly selected subject responds YES to the statement assigned.
- R be the proportion of subjects in the sample who respond YES.

a. (10) Compute the probability r that a randomly selected subject in the population responds YES to the statement assigned. Show that R is an unbiased estimate of r. Take into account that the assignment of statements is independent of the selection procedure.

Solution: The probability that a randomly selected subject in the has property A is q. The assignment of statements is random and independent of the sampling procedure which means that the subject is assigned statement (1) with probability p independently from whether it has property A or not. Since we assume that the subjects give an honest answer, the conditional probability that the subject responds YES given that she is assigned statement (i) is q. Similarly if the subject is assigned statement (2) the conditional probability the she will repond YES is $1-q$. By the law of total probabilities we have that $r = pq+(1-p)(1-q) = 1-p-q+2pq$.

To see that R is an unbiased estimator of r write $R = \frac{1}{n}\sum_{i=1}^{n} I_i$where I_i is the indicator of the event that the i-th subject in the sample responds YES. Above we have shown that $E(I_i) = P(I_i = 1) = r$ which proves the claim.

b. (5) Suggest an unbiased estimator of q. When is this possible? Express the variance of the estimator with $\mathrm{var}(R)$.

Solution: We have

$$q = \frac{p+r-1}{2p-1}\,,$$

which suggests that

$$Q = \frac{p + R - 1}{2p - 1}$$

is a plausible choice. We have shown that R is an unbiased estimator of r. By linearity of expectation Q is unbiased. The above is only possible if $p \neq 1/2$. If $p = 1/2$ we note that a given subject will repond YES with probability $1/2$ irrespective of the value of q. Whatever expression we take its expectation will not depend on q which means that it cannot be an unbiased estimator. For the variance we compute

$$\mathrm{var}(Q) = \frac{\mathrm{var}(R)}{(2p-1)^2}\,.$$

c. (10) Let N_Y be the number of sample subjects who respond YES to the question and N_A na number of sample subjects who have property A. Assume as known that

$$\mathrm{var}(N_Y) = np(1-p) + (2p-1)^2 \mathrm{var}\left(N_A\right)\,.$$

Compute $\mathrm{var}(R)$. Use this to give the standard error of the unbiased estimator of q.

Hint: what quantity does N_A/n estimate?

Solution: By formulae for simple random sampling we have

$$\mathrm{var}(n_A) = \frac{N-n}{N-1}\, nq(1-q)\,,$$

Taking into account that $R = n_Y/n$ we finally get

$$\mathrm{var}(R) = \frac{p(1-p)}{n} + \frac{N-n}{N-1}\,\frac{(2p-1)^2 q(1-q)}{n}\,.$$

From this the standard error is derived easily.

2. (25) Assume the data $x_1, x_2, \ldots, x_n$ are an i.i.d. sample from the distribution given by

$$P(X_1 = x) = \binom{2x}{x} \frac{\beta^x}{4^x (1+\beta)^{x+\frac{1}{2}}}$$

for $x = 0, 1, \ldots$ and $\beta > 0$.

a. (5) Find the maximum likelihood estimator for the parameter β.

Solution: The log-likelihood function is given by

$$\ell(\beta|\mathbf{x}) = \sum_{k=1}^{n} \log \binom{2x_k}{x_k} + \log\beta \sum_{k=1}^{n} x_k - \log 4 \sum_{k=1}^{n} - \log(1+\beta) \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) .$$

Taking derivatives and equating with 0 *we get the equation*

$$\frac{1}{\beta} \sum_{k=1}^{n} x_k - \frac{1}{1+\beta} \sum_{k=1}^{n} \left(x_k + \frac{1}{2}\right) = 0 .$$

Hence

$$\hat{\beta} = \frac{2\sum_{k=1}^{n} x_k}{n} .$$

b. (5) Convince yourself that

$$\begin{aligned} E(X_1) &= \sum_{k=0}^{\infty} kP(X_1 = k) \\ &= \frac{2\beta}{4(1+\beta)} \sum_{k=1}^{\infty} [2(k-1)+1] P(X_1 = k-1) \\ &= \frac{2\beta}{4(1+\beta)} 2E(X_1) + \frac{2\beta}{4(1+\beta)} . \end{aligned}$$

Use this to show that the maximum likelihood estimator is unbiased.

Solution: The equality can be checked by a straightforward computation. The equality transforms into

$$E(X_1) = \frac{\beta}{1+\beta} E(X_1) + \frac{\beta}{2(1+\beta)}$$

or

$$E(X_1) = \frac{\beta}{2} .$$

We have

$$E\left(\hat{\beta}\right) = E\left(\frac{2\sum_{k=1}^{n} X_k}{n}\right) = \beta ,$$

hence the estimator is unbiased.

c. (5) Use the Fisher information to give an approximate standard error for the maximum likelihood estimator.

Solution: Compute for $n = 1$:

$$\ell'' = -\frac{k}{\beta^2} + \frac{k + \frac{1}{2}}{(1+\beta)^2}\,,.$$

hence

$$E\left(-\ell''\right) = \frac{1}{2\beta} + \frac{\frac{\beta}{2} + \frac{1}{2}}{(1+\beta)^2} = \frac{1}{2} \cdot \frac{1}{\beta(\beta+1)}\,.$$

It follows that

$$\hat{\text{se}}(\hat{\beta}) = \frac{\sqrt{2\beta(1+\beta)}}{\sqrt{n}}\,.$$

d. (10) Convince yourself that

$$\begin{aligned} E(X_1^2) &= \sum_{k=0}^{\infty} k^2 P(X_1 = k) \\ &= \frac{\beta}{4(1+\beta)} \sum_{k=1}^{\infty} \left[4(k-1)^2 + 6(k-1) + 2\right] P(X_1 = k-1) \\ &= \frac{\beta}{4(1+\beta)} \left(4E(X_1^2) + 6E(X_1) + 2\right)\,. \end{aligned}$$

Compute the exact standard error of the maximum likelihood estimator.

Solution: The equality is checked by a straightforward caclulation. We get the equation

$$E(X_1^2)(1+\beta) = \beta E(X_1^2) + \frac{3\beta}{2} E(X_1) + \frac{\beta}{2}$$

or

$$E(X_1^2) = \frac{\beta(2+3\beta)}{4}$$

and as a consequence

$$\text{var}\left(X_1\right) = \frac{\beta(1+\beta)}{2}\,.$$

the exact variance of the estimator $\hat{\beta}$ is

$$\text{var}(\hat{\beta}) = \frac{2\beta(1+\beta)}{n^2}$$

3. (25) Gauss' gamma distribution is given by the density

$$f(x,y)=\sqrt{\frac{2\lambda}{\pi}}\,y\,e^{-y}\,e^{-\frac{\lambda y(x-\mu)^2}{2}}\,.$$

for $-\infty < x < \infty$ and $y > 0$ and $(\mu,\lambda) \in \mathbb{R}\times(0,\infty)$. Assume that the observations are pairs $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$ generated as independent random pairs $(X_1,Y_1),\ldots,(X_n,Y_n)$ with density $f(x,y)$. We would like to test

$$H_0\colon \mu=0 \qquad \text{versus} \qquad H_1\colon \mu\neq 0\,.$$

a. (15) Compute the maximum likelihood estimates of the parameters. Compute the maximum likelihood estimate of λ when $\mu=0$.

Solution: The log-likelihood function is

$$\ell=\frac{n}{2}\log\left(\frac{2\lambda}{\pi}\right)+\sum_{k=1}^{n}(\log y_k-y_k)-\frac{\lambda}{2}\sum_{k=1}^{n}y_k(x_k-\mu)^2\,.$$

Equate the partial derivatives with 0 to get

$$\frac{n}{2\lambda}-\frac{1}{2}\sum_{k=1}^{n}y_k(x_k-\mu)^2=0$$

and

$$\lambda\sum_{k=1}^{n}y_k(x_k-\mu)=0\,.$$

From the second equation we get

$$\hat{\mu}=\frac{\sum_{k=1}^{n}x_ky_k}{\sum_{k=1}^{n}y_k}\,.$$

Insert this into the first equation to get

$$\hat{\lambda}=\frac{n}{\sum_{k=1}^{n}y_k(x_k-\hat{\mu})^2}\,.$$

When $\mu=0$ the first equation determines λ. We get

$$\tilde{\lambda}=\frac{n}{\sum_{k=1}^{n}x_k^2y_k}\,.$$

b. (10) Find the likelihood ratio statistics for the above testing problem. What is its approximate distribution under H_0?

Solution: Ths test statistic is

$$\begin{aligned}\lambda &= 2\Big[\ell(\hat{\lambda},\hat{\mu}|\mathbf{x},\mathbf{y}) - \ell(\tilde{\lambda},0|\mathbf{x},\mathbf{y})\Big] \\ &= n(\log\hat{\lambda} - \log\tilde{\lambda}) - \hat{\lambda}\sum_{k=1}^{n} y_k(x_k-\hat{\mu})^2 + \tilde{\lambda}\sum_{k=1}^{n} x_k^2 y_k\,.\end{aligned}$$

However, from the equations for estimates we get that

$$\hat{\lambda}\sum_{k=1}^{n} y_k(x_k-\hat{\mu})^2 = \tilde{\lambda}\sum_{k=1}^{n} x_k^2 y_k = n\,,$$

which implies

$$\lambda = n\log\frac{\hat{\lambda}}{\tilde{\lambda}}\,.$$

By Wilks's theorem, under H_0 the distribution of the test statistic is approximately $\chi^2(1)$.

4. (25) Assume the regression model

$$Y_k = \beta x_k + \epsilon_k$$

for $k = 1, 2, \ldots, n$ where $\epsilon_1, \ldots, \epsilon_n$ are uncorrelated, $E(\epsilon_k) = 0$ and $\mathrm{var}(\epsilon_k) = \sigma^2$ for $k = 1, 2, \ldots, n$. Assume that $x_k > 0$ for all $k = 1, 2, \ldots, n$. Consider the following linear estimators of β:

$$\begin{aligned}
\hat\beta_1 &= \frac{\sum_{k=1}^n x_k Y_k}{\sum_{k=1}^n x_k^2} \\
\hat\beta_2 &= \frac{1}{n}\sum_{k=1}^n \frac{Y_k}{x_k} \\
\hat\beta_3 &= \frac{\sum_{k=1}^n Y_k}{\sum_{k=1}^n x_k}
\end{aligned}$$

a. (5) Are all estimators unbiased?

Solution: Since $E(Y_k) = \beta x_k$ for all $k = 1, 2, \ldots, n$ all the estimators are unbiased.

b. (10) Which of the estimators has the smallest standard error? Justify your answer.

Solution: All the estimators are unbiased. Guass-Markov tells us that the best estimator is the one given by least squares and that is $\hat\beta_1$.

c. (10) Write down the standard errors for all three estimators.

Solution: We first compute the theoretical variances. Since $Y_1, \ldots, Y_n$ are uncorrelated we have

$$\begin{aligned}
\mathrm{var}(\hat\beta_1) &= \frac{\sigma^2}{\sum_{k=1}^n x_k^2} \\
\mathrm{var}(\hat\beta_2) &= \frac{\sigma^2 \sum_{k=1}^n x_k^{-2}}{n^2} \\
\mathrm{var}(\hat\beta_3) &= \frac{n\sigma^2}{\left(\sum_{k=1}^n x_k\right)^2} .
\end{aligned}$$

We need an unbiased estimate of σ^2. Theoretically we have that

$$\hat\sigma^2 = \frac{1}{n-1}\sum_{k=1}^n (Y_k - \hat\beta x_k)^2$$

is an unbiased estimator σ^2. This gives us an unbiased estimator of σ^2.